TERM OF REFERENCES

Term of References (TOR) contains requirements and works that must be done by the student during the internship program for fulfilling the research projects that will be conducted at Deguchi Laboratory in National Institute of Technology, Gifu College (NIT, Gifu College).

1. Objectives

Generally, this internship program aims to develop basic research skill of the student in some specific fields that depends on their research topic and which laboratory will the student be part of. In addition, this internship program is expected to be a bridge of the two countries, Indonesia and Japan, in order to maintain a good relation, especially between Bandung Institute of Technology (ITB) and NIT, Gifu College. This program is also expected to improve the students' communication skill from both of the schools through foreign languages.

On the research implementation, the student will be given several research topics that belong to several laboratories and thereafter the student has to choose their preferred topic. Later, the student will do some research projects based on the selected topic. The research projects will be conducted and supervised by a professor of the laboratory for about 5-6 weeks. At the end of the internship program, the student shall submit an internship report to the internship supervisor in NIT, Gifu College.

2. Project Scope

The research that will be conducted during the internship program is related to computer science field of study, specifically text mining. Furthermore, the research will be held at Deguchi Laboratory and supervised by Prof. Toshinori Deguchi. The student will carry out total two projects, which are:

a. Word Similarity Measure using WordNet

   Within this project, the student should be able to make a program that receives total two words as input. After that, program should be able to give numbers that represent how similar the two given words are. The calculation itself is based on Wu and Palmer similarity measure. For further details, after constructing the conceptual similarity tree of the two given words, Wu and Palmer similarity measure will be performed based on counting the distance between root, least common super-concept of the two words, and the two words themselves. WordNet SQLite will be used within the project and acts as the database for finding the words and the linkages between them.

b. Document Similarity Measure

As more advanced compared to the previous project, the student should be able to calculate similarity in a larger scope that is document. A program for calculating the similarity between documents must be delivered as one of the results of this project. The calculation itself is divided into two methodologies, which are using vector space model or sometimes called as term vector model and using latent semantic analysis (LSA). For both methodologies, there are several term weighting functions that will be used, which are binary, term frequency (tf), and term frequency-inverse document frequency (tf-idf).

3. Project Results

The output of this internship program can be classified as two groups, which are from the first project and the second project. The results from each of them are similar, a program for calculating the similarity between words will be delivered from the first project and a program for calculating the similarity between documents will be delivered from the second project. For the two projects, a final report in the form of document must also be delivered at the end of the internship program.

4. Project Constraints

The followings are some constraints that are applied throughout the first project about Word Similarity Measure using WordNet:

1. Program is developed using WordNet SQLite.
2. Program is written in C/C++ language.
3. Similarity calculation is based on Wu and Palmer similarity measure.
4. Similarity calculation is applied only to noun (n).
5. The similarity tree is constructed based on searching the hypernyms of each synsets of the given words.

The followings are some constraints that are applied throughout the second project about Document Similarity Measure:

1. Program is written in R language.
2. Similarity calculation are based on cosine similarity from the vector space model and latent semantic analysis.
3. Three kind of term weighting functions is used for constructing the vector space model, which are binary, term frequency (tf), and term frequency-inverse document frequency (tf-idf).
4. Book summaries dataset is used within the project.

5. For testing and validating the program, the dataset is focused on books which have sequel.

5. Assumptions

The followings are some assumptions regarding status or conditions which are agreed for conducting the Word Similarity Measure using WordNet project:

1. The student has at least basic knowledge about computer science field of study and some programming languages such as C/C++.
2. The student eagers to learn about WordNet and SQLite especially SQLite library for C/C++ for creating the program.
3. The research project can be done within 2 weeks.

The followings are some assumptions regarding status or conditions which are agreed for conducting the Document Similarity Measure:

1. The student has at least basic knowledge about computer science field of study and some programming languages such as C/C++.
2. The student eagers to learn R language and some methods for calculating the document similarity.
3. The research project can be done within 3 weeks.

6. Work Methodology

In order to carry out the research project, there are some strategies, including the method, process, or tools, which are summarized in a work methodology of the projects and enlisted below:

Project 1	1. Set up work environment for the first project. Tool, libraries, and other dependencies installation are included. 2. Learn about SQLite and database scheme of WordNet. 3. Learn about how to use WordNet in C/C++. 4. Learn about similarity between words. 5. Program development. 6. Program testing. 7. Make report.
Project 2	8. Set up work environment for the second project. Tool, libraries, and other dependencies installation are included. 9. Learn about document similarity in general. 10. Learn about vector space model and some term weighting functions. 11. Learn about cosine similarity. 12. Program development (first iteration).

	13. Program testing (first iteration). 14. Learn about eigenvalue, eigenvector, and singular value decomposition (SVD). 15. Learn about latent semantic analysis. 16. Program development (second iteration). 17. Program testing (second iteration). 18. Make report.
Final	19. Final report.

7. Environment Development

The following information is the environment details in developing the program for Word Similarity using WordNet project:

a. Hardware Specifications

Operating system : Windows 10 Pro
Processor : Intel® Core™ i3-3110M CPU @ 2.40GHz
RAM : 8.00 GB (7.89 GB usable)
System type : 64-bit Operating System, x64-based processor

b. Software Specifications

Project 1	DBMS	: SQLite ver. 3.8.10.2
	Language	: C++, PHP, HTML, CSS, JavaScript
	IDE	: Code::Blocks 16.01
	Backend/server	: Apache 2.0 Handler
Project 2	Language	: R

8. Schedule

Generally, the student must come to the laboratory every Monday-Friday, from 9.00 a.m. until 16.10 p.m. for carrying out the research project. Despite that, other activities might be applied during the research time e.g. attending a welcoming party, giving presentation in some classes, etc. The detailed schedule can be seen in Figure 1 below.

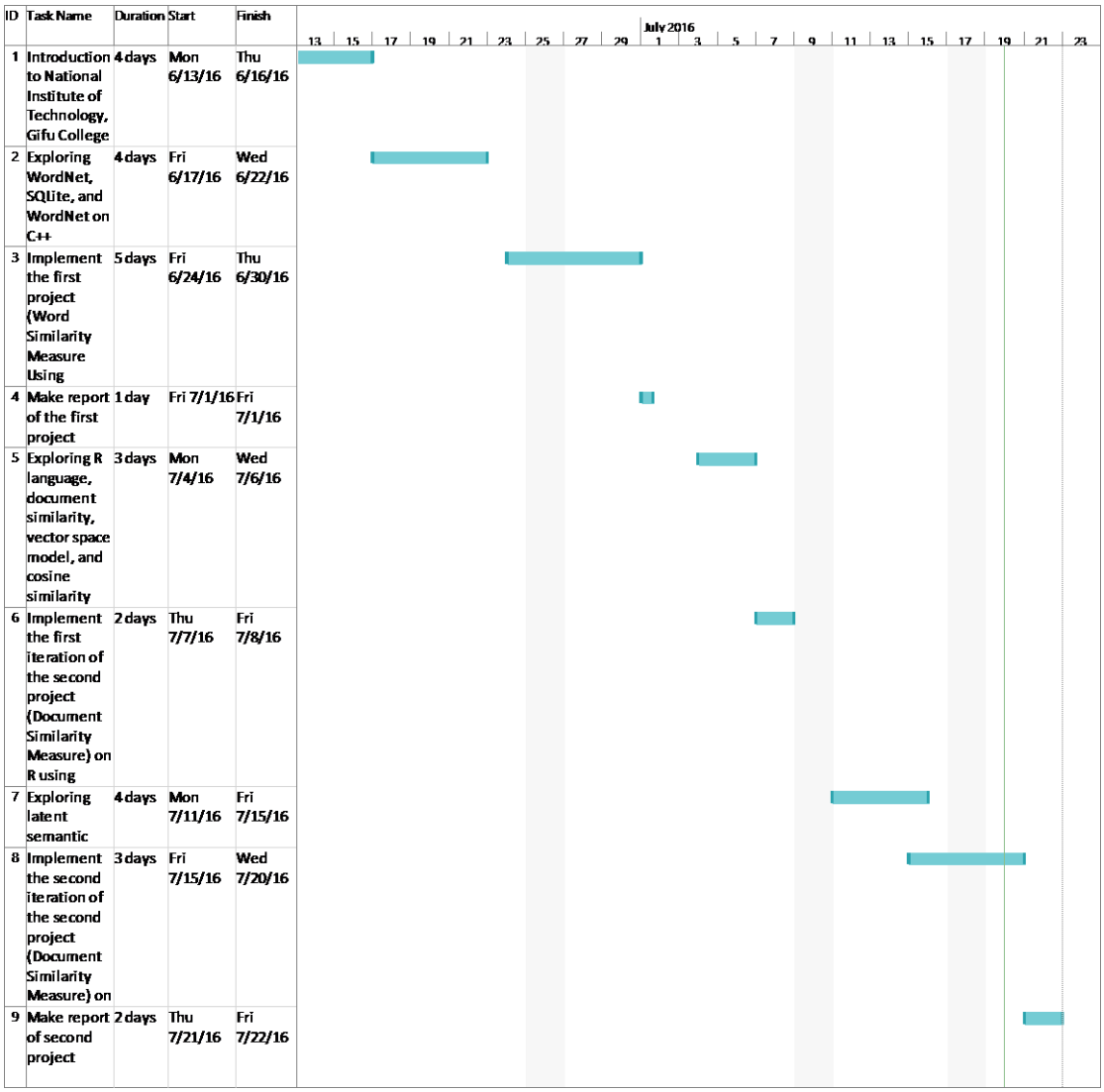

Figure 1. Internship program Gantt chart

9. Staffing

No staffing needed for this internship because the student will carry out the research by oneself which means no other student will be involved.

10. Control and Monitoring Mechanism

During participating in the internship program, especially when carrying out the research projects, the student will be monitored by a professor of the laboratory that the student will be part of. The professor also acts as a supervisor for the projects. The followings are further details of control and monitoring mechanism for the internship program:

1. Lab activities begins from 09.00 a.m. until 16.10 p.m. for every weekdays (Monday-Friday). The student will be monitored by the internship supervisor at NIT, Gifu College during working in her projects.
2. There are several milestones for the project works, enlisted below:

a. June 22, 2016 : The student has completed doing some explorations and learning all materials needed for the first project.
b. July 1, 2016 : The student has successfully made all of first project deliverables, which are word similarity measure program, document report, and presentation.
c. July 8, 2016 : The student has successfully made the first iteration of document similarity measure program.
d. July 20, 2016 : The student has successfully made the second iteration of document similarity measure program.
e. July 22, 2016 : All results from both projects have been delivered.

3. At the end of the internship program, the student will give a final report about the projects.

11. Other Information

Student will receive JASSO scholarship during the internship program for sponsoring student's daily expenses in Japan.

Approved by,

Prof. Toshinori Deguchi	Tifani Warnita
Internship Supervisor	Internship Student